On lexical indices. Illustrations from Italian texts

Jane Nystedt

1. Introduction.

Within the field of stylolinguistics there has long been a discussion of the so-called lexical richness of a given literary text, how to describe it and how to measure it. Linguists as well as mathematicians have elaborated various indices, trying to find ways of measuring such lexical richness. We find experiments in this way in Herdan (1966), Guiraud (1954), Muller (1968), Dugast (1978) and in Brunet (1978), who have all reelaborated and come back to varieties of their indices in later works.

The need for such a measure becomes apparent when literary texts are compared since an important problem encountered is that the compared texts usually are of different lengths. Therefore, some kind of index is needed which neutralizes this difference. The simplest index would be just to count the number of different forms, V, in the text. This is clearly not good, since V naturally increases with the number of words, so that a longer text automatically receives a higher index than a shorter text. An easy way, then, would be to compute V/N; however, it soon becomes evident that this 'overcompensates' for N so that the quotient always becomes smaller as N gets larger.

The search for a 'fair' index has resulted in several suggestions of indices with more or less complicated formulas involving N and V. The results of these formulas are numbers which are supposed to measure the lexical richness, independently of N. The numbers such indices give are abstract figures which are not easy to interpret. Moreover, until now it has also appeared impossible to compare one index with another, at least it has never been done. This will, however, be accomplished in this paper with particular reference being given to the works in prose by the

(ricevuto nel marzo 1993)

Italian author Primo Levi.[1] I wanted to compare the different texts of Levi from the viewpoint of lexical richness but did not know if the methods in use were partial to either long or short texts. I also found the abstract numbers, which were the results of these methods, difficult to interpret. I began to experiment with the various indices on excerpts from Levi texts of well-defined lengths and then tried to transcribe the results in a new way which will be explained below.

In this paper we shall discuss the concept of lexical richness and some of the possible indices used to measure this richness. We shall also develop two methods to make such indices more understandable and less abstract. Further, we will use these methods in order to compare different indices and determine if a certain index is a 'fair' index, i.e. if it treats shorter and longer texts equitably or if it favours texts of a certain length.

2. *Postulates.*

Let us postulate the following:

1 - There exists something called the 'lexical richness' of a text; it is not yet quite clear what this is exactly, but we have an intuitive feeling of it.

2 - The lexical richness can be measured by a lexical index; we do not yet know how to construct such an index.

3 - The lexical richness shall be largely independent of the length of a text, at least within certain bounds, which however shall be quite wide. When a text is extremely short every word is probably a hapax and every new word is a new form. When a corpus is extremely big, say the collected novels of all Italian authors since Dante, it is probable that an additional novel will add only few new forms, if any, to the ones already present.

4 - A literary work is called 'homogeneous' if any part of the book, regardless from where it has been taken, has the same lexical richness as any other part or the whole work. We shall actually use this postulate in reverse; we shall pronounce intuitively that a certain work is 'homogeneous'. We shall also justify this subjective feeling with lexical experimenting, described below. Then we shall judge an index as 'fair' if it has the same value for the entire work as for parts of the work, regardless of their lengths. We shall then say that we have obtained a lexical index which is not sensitive to the length of the text.

[1] My interest in the matter has developed from the stylolinguistic study of the works in prose by Primo Levi which I have recently conducted under the auspices of The National Swedish Research Council. - All the works by Primo Levi referred to in this article are included in *Opere*.

3. *Lexical richness.*

We have an instinctive feeling that childrens' books are not 'lexically rich': they contain no foreign or 'difficult' words, hence their vocabulary is smaller. Scientific papers are also often 'lexically poor': although they have a highly specialized vocabulary, it is, in general, quite limited and the rest of the text is usually rather poor as well. We could say that a text is lexically rich if we feel that the author uses a varied language, has many synonyms and tries to avoid repetitions etc. This article will not discuss the concept of a 'literary index'; we intuitively know that it is not the same as the idea of the 'lexical index' which is treated here. We will leave the development of a way to measure a 'literary index' to others, should they consider this meaningful.

4. *The writing of a book, according to statistical linguistics.*

'The theoretical model of writing' is quite remarkable. It may be described as follows. The writer has by her side a large urn filled with slips of paper. On each slip is written a word. Some words, e.g. 'and', 'or', 'you', are written on numerous slips, whereas other words, such as 'autonomous' or 'semipelagianism', are much more rare. The process of writing consists in picking slips, one after another, from the urn and copying the words written on each slip. When the writer has drawn and copied enough slips, she stops and the article, novel or play is completed.

Thus a story may begin: "moment whenever of out enemy drink umbrella ...". It is amazing that such an unrealistic model can produce results which may be confirmed by experiments.[2] Notice that a text composed 'at random' in this way may be 'lexically rich' but its literary value is definitely nil.

The totality of different slips in the urn may be called the 'theoretical lexical richness' of the text. If the urn contains many different words, the lexical richness is greater than if the urn contains only a few different words. Hence, it is quite clear that a word that is not in the urn cannot appear in the text since it cannot be drawn. On the other hand, a word might very well be part of the urn content, yet still not appear in the text. It is quite enough that it could have been included and as such of great importance.[3]

The content of the urn varies according to the intentions of the writer. It is not the same when she writes an essay on statistical linguistics as when she writes a letter to her husband. In fact, it varies with every new word she puts on the paper (or in the word processor). If the two first words of a sentence are "I am ...", the urn cannot contain words

[2] See e.g. Guiraud (1959), Muller (1973).

[3] The urn-model has been thouroughly discussed by Guiraud (1959) and Muller (1968).

such as 'are', 'have' or 'girl', whereas slips with 'a' and 'an' are plentiful. If the next word is 'a', the content of the urn changes. The slips with 'a' and 'an' vanish and instead other slips such as 'girl' perhaps come forth. However, these short-term changes cancel out one another, and we can consider the urn to be constant for several pages at least.

When the story of a novel or a play evolves there is a change of scene, the events are new and this means that new words are added to the urn and other words disappear. There may be pages filled with dialogue in direct citation and for these pages the content of the urn probably becomes much more meager. All these changes are weighed together, and the result is this abstract nebulous urn called lexical richness.

In the urn model, the possibility for a particular word to be chosen is K/L, where L is the total number of slips in the urn and K is the number of slips with this particular word written on it. The sum of possibilities of all the words is of course = one. The urn model is not abstract enough: there is no possibility between 1/L and 2/L and no positive possibility smaller than 1/L. We need a model where a word can be written on half a slip or on 19/12 of a slip, i.e. a continuous model, where the possibility of a word to be chooseh can be any positive number less than one; the sum of all possibilities is, as said above, by definition one. This would imply that if the number of different words in the urn becomes greater, the possibilities for the words already existing in the urn must diminish in order to make room for the new words.

We would like to judge the lexical richness from the content of the urn: the more different word forms in the urn, the greater the lexical richness. To us, it does not matter if the words in the urn are common and easily understood or if they are foreign and unusual, it is only the amount that matters. Of course, the writer's urn - the theoretical lexical richness - is unaccessible to us. We only have the text itself with which to work.[4]

5. *The lexical index.*

The common procedure to compute a lexical index in order to measure the lexical richness of a given text is to make a computation from the number of occurrences, words, N, and the number of different word forms, V.

The interest of a lexical index lies in its ability to measure 'the lexical richness'. Given two texts, one with N_1 words, occurrences, and V_1 different forms, and the other with N_2 words and V_2 forms, we compute the lexical indices for the two texts. From this we decide which text is richer or that the two texts are equally rich. An index must be able to

[4] Some scholars have tried to estimate the extent of the urn from the literary work, e.g. the study of Efron & Thisted (1976) on the theoretic vocabulary of Shakespeare.

compare texts of which one could be ten times the length, or more, of another.

Indices of this kind were already devised, as stated above, by linguists as well as mathematicians. Later, amongst others, Brunet (1985) further developed and used his index W, Dugast (1979) index U, which, according to his article of 1978, is "une adaptation" of Guiraud's index. Very few of these authors do however discuss other indices than their own and no one makes experiments in order to compare them in practical measuring of the lexical richness of a text. Except for Dugast (1978), only Bernet (1983) dedicates somewhat abundant space to a description of the indices so far in use and Brunet (1985) merely indicates the formulas of Guiraud's and Dugast's indices.

We shall now discuss index W, according to the formula [5]

$$W = N^{V^{-0.185}} \quad (1)$$

and index U, according to the formula

$$U = \frac{\ln^2 N}{\ln N - \ln V} \quad (2)$$

On the basis of numerous experiments with Levi-texts, we shall investigate these two already known indices W and U, as well as a new index, H, to see if, and to what extent, they are impartial to the length of a text. We shall also try to make the indices more understandable, so that they convey an immediate meaning to us.

6. *The homogeneity of a text.*

Postulate 4 above pronounced a literary work as 'homogeneous' if any part of the text has the same lexical richness as any other part or the whole book, regardless of the size of the excerpt.

This definition raises the problem of how to decide whether a certain text is homogeneous or not. The immediate answer, namely that this can be measured by means of a lexical index, is of course not conclusive until we know that the index used is a 'fair' one, that is to say that it is impartial to the length of the excerpt. Apparently, at this point, we must do something to avoid a circular argument.

[5] In Brunet (1978:49), instead of -0.185, the author uses -0.172 and in Brunet (1983:19) -0.134; it is not clear why he, in his later studies, changes the formula.

Suppose we apply two different indexes to a certain text and obtain divergent results. One index says that the text is homogeneous, and the other indicates that it is not - what shall we then believe?

What we need is a set of texts of various lexical richness, all of which we know by divine revelation that they are homogeneous. For each text T we determine the corresponding curve $V = f_T(N)$, where the subscript T indicates the text T from which the curve is computed.

Now, any lexical index, I, is a one-parametric set of level curves

$$I(N,V) = C$$

in the N,V-plane, where the constant, C, is the parameter (or value of the index). For each value of C, the corresponding curve joins together all pairs (N,V) which give that lexical value, C.

We then say that that index is best, whose level curves best approximate the curves $V = f_T(N)$ for all T. (There might even be no such index!) When all this is done, this index I may be used for measuring.

The problem is of course that we have no God-given standard, in the name of which we can pick the homogeneous texts. The best we can do, is to discard texts which for some reason (as for instance is the case with some of the experiments below) we suspect are not homogeneous, and hope that we are left with at least acceptably homogeneous texts.

Hopefully, experience will then teach us which index we can confide in, and in which we cannot. It may not be the same index for every language, nor indeed for every, say, Italian author.

In order to discuss homogeneity, let us consider the following examples (which, for the sake of arguing, are choosen somewhat on the extreme side). All texts are 40.000 words long. Text A is a 40.000 word essay on music. Text B is also an essay on music, but the first half of the text, 20.000 words, is repeated word for word, so that the latter half is totally identical with the first one. In text C, the first half is the same as in text B, but the second part is a translation of the first into another language and thus the two parts have no words in common. Text D is also an essay of which the first half is on music and the second on chemistry. Finally we have text E, which is also an essay on music but where the first half is written in a very rich language for an exclusive audience, whereas the second part is written in a language suited for pupils in a primary school.

Of these five examples we believe that text A stands the best chance of being homogeneous. Text B ought to have a much lower lexical richness for the whole text than for its first half. Text C, on the contrary, has an abnormally high richness for the whole text compared to its first half.

What can we expect from text D compared to text A? Let us suppose that the two halves of D have the same lexical richness. In the latter part

of text A there is a good chance that words like 'Bach', 'sonata' or 'pentatonic' which are found in the first half will be repeated and that chemical words like 'molecules', 'osmosis' or 'catalyst', not found in the first half, will not appear in the second half either. In text D the reverse is probable. Subjectively we feel that text D is not homogeneous and that, although the first and the last half are equally rich, the whole of D is richer than each of its two halves. We also believe that text D contains more hapaxes than text A. For, we think, obvious reasons we feel that neither text E is homogeneous.

Examples D and E have been choosen to show that a change in the conceptual or narrative level does have consequences on the lexical level as well. We also sense that in text E the second half is lexically poorer than the first half, although this is something that cannot be taken for granted.

Do we have a possibility to detect the non-homogeneity in texts B-E by means of word-counting? Let us first take excerpts of texts A to E, all 20.000 words long, namely the first half and the latter half of each text. Since all excerpts are of the same length, there is no problem of index. We just compare the different values of V. In this way, we reveal that text E is not homogeneous, but text B and C give no indication of anomaly.

If, however, we also choose words 1 - 10.000 and 20.001 - 30.000 as one excerpt, texts B and C will immediately be spotted as non-homogeneous when we compare this excerpt with words 1 - 20.000. The same might be expected for text D.

Thus, suppose that, from a certain text T, we compose several excerpts of length N, choosen not only as one continuous 'chunk', but also composed of smaller parts taken here and there. Let us furthermore suppose that we find these excerpts to be equally rich (within tolerable bounds), and that this is the case for a large range of values of N. We then believe it reasonable to think that the text is homogeneous.

Experiences from many texts will then teach us if one index is better than another.

7. *The experiment.*

For this experiment we have used the Oxford Concordance Program, OCP,[6] the Micro-OCP version, allowing it to work on unlemmatized text.[7] Our main object was to try out our ideas on a text which was supposed to be lexically homogeneous. Therefore we chose to let OCP

[6] The Oxford Concordance Program was elaborated by Susan Hockey, Oxford, and is distributed by Oxford University Press.

[7] All studies so far have been made on unlemmatized text, partly due to the fact that lemmatization has been considered an extremely time-consuming and tideous task because of the lack of qualified and suitable software. Further, it is considered that studies on lemmatized text would reveal less about form and structure but more about contents.

work on excerpts from Levi's *Se non ora, quando?* (hence referred to as *SNOQ*), from *La tregua* (hence referred to as *T*) and also from one of the chapters of *I sommersi e i salvati* (hence referred to as *SES*). We also experimented on other texts by Levi which we did not consider to be homogeneous.

The primary reason we chose *SNOQ* and *T* was because they consist of one single story: they are not collections of disparate essays or short stories. Secondly, we had the subjective impression that they are rather homogeneous from the point of 'lexical richness' since their stories develop evenly, there are no abrupt changes of style etc. We therefore postulate, as our point of departure, that excerpts from these books have the same 'lexical richness' as the whole book, independent from where these excerpts are taken. The same applies to the chapter chosen from *SES*.

With the aid of OCP, we can study a complete text as well as parts of a given text. These parts may be taken in one 'chunk', e.g. words number 10001 to 20000, or we may pick them in several smaller pieces, e.g. words number 12001 to 17000 and 62001 to 67000.[8]

For our experiment we took samples from *SNOQ* of various lengths N, where N was respectively 78, 156, 313, 626, 1250, 2500, 5000, 10000, 20000, 40000, 80000, 100000 words and, finally, the whole of *SNOQ*, 109598 words.[9] The ratio between the whole of *SNOQ* and the shortest excerpt is over 1400. It may be discussed how long an excerpt must be before it is meaningful to talk about its lexical richness. It might be argued that 500 words is a minimum. (It should be pointed out that in our experiment, we went as low as N = 69 words when *T* was studied, and N = 78 when *SNOQ* was studied.)

It can be noticed that each value of N is twice the former value, up to 80000. For each sample we let OCP count V, i. e. the number of different word forms. Homonyms were not separated and e.g. each of the different forms of a verb was counted separately.[10] OCP also counted many other things such as the number of hapax legomenon and hapax dislegomenon.[11] The result of these counts have been discussed in Nystedt (1993).

From *SNOQ* we took several samples of each N. For N = 20000 words, for instance, we took 9 samples, such as shown in table 1: words 1-20000, words 20001-40000, words 400001-60000, 60001-80000 and 80000-100000. We also took samples such as words (1-5000) + (20001-25000) + (40001-45000) + (60001-65000). The reasons for

[8] Several scholars, in experimenting with lexical indices, have divided a given text in excerpts, however always in samples of equal length. Often quoted in this respect is Muller (1971, 1977).

[9] In Nystedt (1993), all data have been computed with the Italian program DBT, elaborated at the Istituto di Linguistica Computazionale, Pisa, Italy, see Picchi (1989). Since DBT and OCP operate differently, the figures for N and V for the various Levi texts are here slightly different.

[10] The Italian originals of Primo Levi's books have probably a higher lexical index than for instance their Swedish translations since modern Swedish has very few different verb forms.

[11] The originally Greek words *hapax legomenon* stand for wordforms that appear only once in a text, *hapax dislegomenon* for words that appear twice in the same text.

our way of sampling was explained above, in the paragraph regarding the homogeneity of a text.

Excerpt No.	*Words number, N*	*V*
1	1-20000	4340
2	20001-40000	4744
3	40001-60000	4715
4	60001-80000	4227
5	80001-100000	4712
6	15001-25000	
	95001-105000	4700
7	1-10000	
	65001-75000	4287
8	20001-30000	
	80001-90000	4690
9	1-5000	
	20001-25000	
	40001-45000	
	60001-65000	4629
	V: mean = 4560	
	std. dev = 198.90	

Table 1. *Samples of N = 20000 words from SNOQ, and number of word forms, V.*

We assumed that a comparison between the samples would reveal any significant difference between the beginning, the middle and the end of *SNOQ*, and also any difference if we took samples part from the beginning, part from the middle of the book. The reason for this was that we suspected that as the story evolved, the vocabulary would change and in this way we hoped to discover if our assumption on the homogeneity of *SNOQ* was erroneous. Considering the first 5 'chunks' of N = 20000, we note that the values differ from 4227 to 4744, a difference of 10% only.

As shown in table 1, the first 20000 words of the book are not lexically richer than later units of 20000 words, which was unanticipated. The reason we expected this is that in the beginning of a book factors like the setting, the characters and to some extent the plot are introduced. Apparently, it does not function that way, at least not with Primo Levi. It wouldn't have been astonishing if the V values for excerpts 6-9, table 1 above, had been higher than the others, since each of the individual parts constituting the excerpts has its own specific vocabulary. This is particularly conspicuous with sample 9, where the words were picked from four different sections of the book and where consequently V could have been expected to be noticeably higher. It

isn't however and, together with the results of samples 6, 7 and 8, it confirms our hypothesis of the homogeneity of *SNOQ*. The standard deviation of V for these nine samples is 199 or 4% of V, a figure we think does not repudiate the homogeneity of *SNOQ*.

In all, we had as many as 80 samples of different sizes from *SNOQ*, see table 2. Each of the lines in this table summarizes the results of a table like table 1. Table 1, in turn, is summarized in the line of table 2, marked off by a '*'. We have not, however, reproduced all those tables here since we did not feel this to be necessary. For each N, column 1, we took the mean and the standard deviation of the number of forms, V, columns 2 and 3. Column 4 gives the standard deviation of V in per cent of V. In the column on the far right we give the number of samples studied for each N.

	Number of words, N	*Forms mean, V*	*Std dev, V*	*Std dev of V in % of V*	*Number of samples*
	78	60	8.56	14%	6
	156	112	6.77	6%	6
	313	195	11.91	6%	6
	626	356	14.06	4%	6
	1250	569	31.62	6%	6
	2500	945	68.73	7%	7
	5000	1675	79.31	5%	12
	10000	2796	147.61	5%	12
*	20000	4560	198.90	4%	9
	40000	7232	235	3%	6
	80000	11180	113	1%	2
	100000	12854	-	-	1
	109598	13466	-	-	1

Table 2. *Protocol from all of the 80 samples from SNOQ.*

As is seen from table 2, the standard deviation of V for almost all values of V from *SNOQ* is 6% or less and, as will be seen from tables 3 and 4, even less for *T* and for *SES*.

We excerpted from *T* in a similar manner, the results from all of the 39 samples shown below in table 3; we did likewise for *SES*, with the results from the 15 different samples as shown in table 4.

Number of words, N	*Forms mean, V*	*Std dev, V*	*Std dev of V in % of V*	*Number of samples*
69	57	3.74	7%	4
136	99	2.69	3%	4
270	185	4.02	2%	4
537	325	9.82	3%	4
1072	590	24.48	4%	4
2141	1010	14.87	1%	4
4280	1783	40.54	2%	4
8556	3103	108.05	3%	4
17112	5025	329	7%	4
34225	8273	431	5%	2
68450	13187	-	-	1

Table 3. *Protocol from all of the 39 samples from T.*

Number of words, N	*Forms mean, V*	*Std dev, V*	*Std dev of V in % of V*	*Number of samples*
608	348	7.43	2%	4
1212	621	18.47	3%	4
2424	1075	18.10	2%	4
4848	1848	17.00	1%	2
9694	3140	-	-	1

Table 4. *Protocol from all of the 15 samples from a story of SES.*

When these figures are visualized in a ln-ln-diagram, we get the following, with the figures from *SNOQ*, taken from table 2 above:

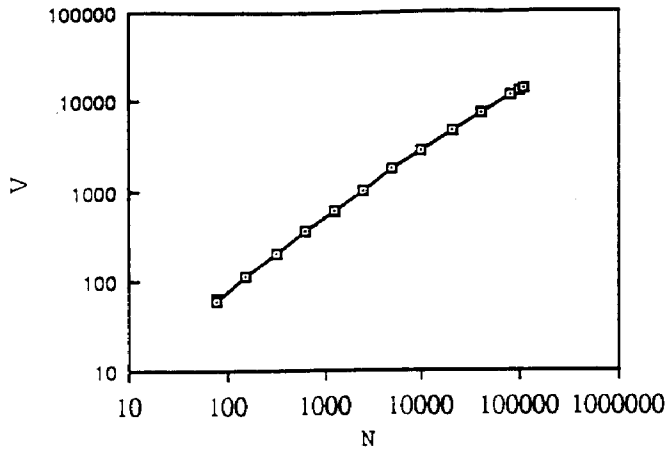

Diagram 1. *Ln-N, ln-V diagram for table 2.*

We see that the points constitute an almost perfect straight line. Standard regression analysis gives the following equation for this line:

(3) $$\ln V = 0.74026 * \ln N + 1.0230$$

with a correlation coefficient of 0.9990 for *SNOQ*, which indeed indicates a very high correlation.

On the basis of this experiment we constructed the following index H: if a certain text contained N words and V forms, we gave it index H, where

(4) $$H = (\ln V - 0.74026 * \ln N - 1.0230) * 100$$

In this way we can compare the lexical richness of a particular text with the 'mean lexical index' of *SNOQ*. By the 'mean lexical index' of *SNOQ* we understand precisely the regression line (4). The factor 100 in formula (4) was introduced to give a text with a certain percentage more forms or a certain percentage less forms than a mean text from *SNOQ* of equal length, index H = +/- this percentage. Of course, index H = 0 is obtained as a mean value for the regression line. The values for N and V of a certain excerpt do not lie exactly on the line and, therefore, receive a value of H which differs from 0, as in table 5.

Index H is of course in a certain sense absurd since it does not pass the origin in the ln-ln-diagram. Any text which is one word long should have one form. Also the derivative

$$\frac{d(\ln V)}{d(\ln N)}$$

ought to be 1 and not 0.740 for N = 1, since most two-word texts have two forms (see the discussion about index U below). On the other hand, it might be argued that it is irrelevant what happens for such extremely short texts as N = 1 or 2.

We calculated the values of the various indices W, U, and H for the exerpts of *SNOQ* as given in table 2 according to formulas (1), (2) and (4) above, and the results are shown in table 5.

N	*V*	*W*	*U*	*H*
78	60	7.711	72.35	-15.36
156	112	8.244	76.96	- 4.25
313	195	8.726	69.78	- 0.25
625	356	8.770	73.64	8.65
1250	569	9.072	64.61	4.24
2500	945	9.050	62.92	3.66
5000	1675	8.645	66.33	9.59
10000	2796	8.345	66.57	9.51
20000	4560	8.036	66.34	7.12
40000	7232	7.748	65.65	1.93
80000	11180	7.482	64.77	- 5.82
100000	12854	7.388	64.61	- 8.38
109598	13466	7.379	64.23	-10.52

Table 5. *Values of indices W, U and H for excerpts from SNOQ.*

8. *Understanding a lexical index.*

It might have been suitable here to give and discuss two additional tables like table 5, one for *T* and one for the chapter from *SES*. We do, however, now wish to discuss the three lexical indices W, U, and H and compare them with each other in order to decide how well each of them measures the 'lexical richness' of a text. We see that index W is somewhere around 7 and 8 whereas U varies from 62 to 77 and H from -15 to 9. The figures have the inconvenience of being abstract and difficult to interpret as actual measures. It also seems impossible to

compare the indices used until we have a method of presenting them in a more intelligible way. We will illustrate such a method in tables 6, 7 and 8. If an index is to be of any use to us, it must be 'understandable'. Of course, prolonged use of an index similar to the ones above makes us familiar with it and we learn to 'read off' a value. An index can, however, be constructed so that its interpretation becomes more or less obvious. Index W, for example, diminishes as the lexical richness increases, which is contrary to what we would naturally expect from this kind of an index.

In order to facilitate the interpretation of an arbitrary index, which we shall call 'I' (where 'I' stands for either index W, index U, index H or, indeed, any index), we propose to do the following: instead of just computing the value of index 'I' when we have a text with N occurrencies and V forms, we shall compute:
the number of forms (V_1) that a text with (N_1 =) 10000 words should have to get the same value of the index 'I'; we shall call this number '*normed I*' (or '*norm I*').

For example: if a text is 156 words long and has 112 forms, index W for this text is 8.244 (see table 5). A text with 10000 words must have 2884 forms to get index W = 8.244, according to formula (1); we call this number 2884 'normed W' (see table 6, the line marked off with a star). If, on the other hand, the lexical richness is measured by index U, the 156 word text with 112 forms gets index 76.96 (see table 5). A text with 10000 words needs 3321 forms to get the same index U = 76.96 according to formula (2). The same procedure with index H gives 2436 forms, according to formula (4). Such actual numbers of forms are much more tangible for us and could lead to a better understanding of how a certain index functions. We can also better compare indices with one another, just as we have done with normed W = 2884 word forms, normed U = 3321 and normed H = 2436 forms for N = 10000.

In the way described above, we shall now, in table 6, let the indices operate on the figures of N, obtained from *SNOQ*, table 5.

	N	*V*	*norm W*	*norm U*	*norm H*
	78	60	3432	3096	2180
*	156	112	2884	3321	2436
	313	195	2498	2965	2533
	625	356	2467	3160	2771
	1250	569	2269	2690	2652
	2500	945	2282	2597	2637
	5000	1675	2557	2784	2798
	10000	2796	2796	2796	2796
	20000	4560	3081	2784	2730
	40000	7232	3389	2747	2592
	80000	11180	3720	2699	2399
	100000	12854	3848	2690	2338
	109598	13466	3862	2669	2288
	Mean value:		3007	2846	2550
	Std. dev :		565.8	212.0	198.0

Table 6. *The number of forms a 10000-word text must contain in order to obtain the same value of index W, U, and H as a text from SNOQ, with N and V specified by the first two columns.*

Table 6 is visualized in diagram 2.

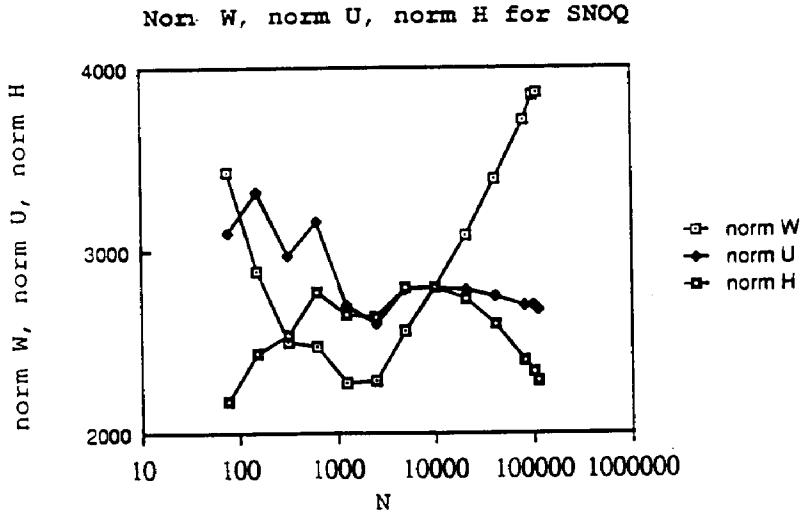

Diagram 2. *Visualization of table 6.*

With the introduction of normed indices, we are now able to discuss the various indices and compare the way they operate on extracts of different lengths from *SNOQ*.

If, as we postulate, *SNOQ* is homogeneous in its lexical richness (which after all is debatable), then a fair index would give approximately the same value for each entry of its column. Apparently this is not the case for any of the three indices as was seen in table 6. A comparison shows a mean value for norm W of 3007 words, for norm U of 2846 words and for norm H of 2550 words. Norm W has values between 2269 and 3862, a span of 1600, norm U lies between 2597 and 3321, which gives a span of 700, and, finally, norm H, which lies between 2180 and 2798, a span of 600. The standard deviation for norm W is 565.8, for norm U, 212.0, and for norm H, 198.0: much higher for norm W than for the other two, the lowest for norm H.

We notice that norm W has a minimum value for N = 1250 where it is 2269 and rises steadily with N to 3862 for the whole of *SNOQ*, an increase of more than 70%. It also increases when N decreases, reaching a value of 3432 for N = 78. The increase is about 50%. The tendency is very clear and without exception. This means that W is partial and favours longer excerpts (>20000 words) as well as extremely short ones, while it is negative towards medium excerpts.

For norm U the minimum value of 2579 is reached for N = 2500, with no definite tendency as N increases. As N decreases, norm U increases, but much less than norm W. The increase is only about 20%. We can draw the conclusion that U is partial to extremely short texts, although less than W. As noticed earlier, it could be discussed if it is of any importance whatsoever to consider the lexical richness in texts shorter than, say, 500 words. It will also be shown below that U is very sensitive to changes in V when N is low.

In view of the almost perfect regression line in diagram 1, we can have rather high hopes for index H, that it shall serve us well. The lowest value, by far, for index H is obtained with texts of only 78 words. Strangely enough, there is also a decline from 100000 words to 109598 words for index H. The values of norm H are 2338 and 2288 respectively. Norm H has a maximum value for N = 5000. The decrease is quite noticeable both for large and small N and amounts to about 18% for the extreme values. This is an unexpectedly high figure in view of the fact that the correlation coefficient for the regression line is 0.9990, as shown above. However, the regression line is computed for ln N and ln V; the logarithm function conceals the magnitude of the differences. Thus, we can say that H favours medium texts but not long or short texts.

Table 6 clearly shows how differently indices W, U, and H behave on *SNOQ*.

A corresponding study of *T* and the chapter from *SES* gave the following results, shown in table 7 for *T* and in table 8 for *SES*.

N	*V*	*norm W*	*norm U*	*norm H*
69	57	3804	4049	2268
136	99	2944	3217	2372
270	185	2725	3581	2678
537	325	2560	3397	2829
1072	590	2647	3533	3081
2141	1010	2717	3381	3160
4280	1783	3007	3457	3342
8556	3103	3403	3501	3394
17112	5025	3699	3349	3376
34225	8273	4200	3312	3327
68450	13187	4730	3240	3175
Mean value:		3312	3456	3000
Std. dev :		683.1	217.5	386.9

Table 7. *The same as table 6 except that the text studied is T.*

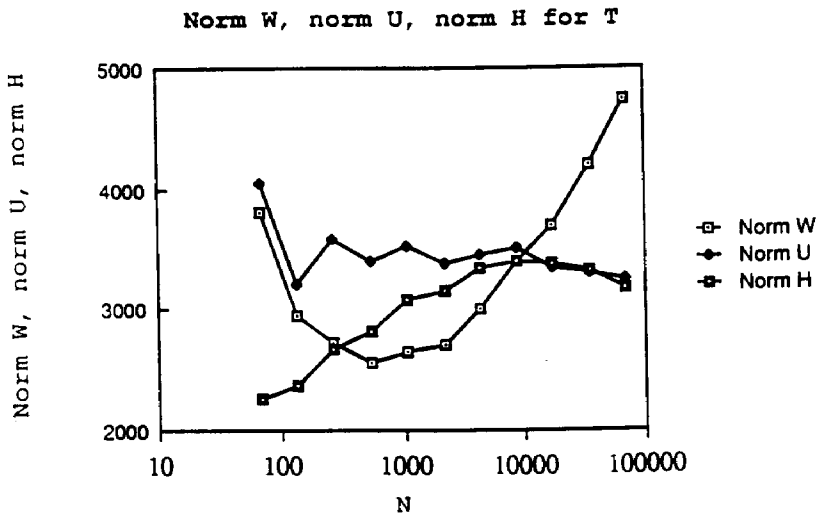

Diagram 3. *Visualization of table 7.*

Normed W varies from 2560 to 4730, a span of 2200, with a mean value of 3312, and a standard deviation of 683.1. It tends to give higher values for very short texts and for long texts.

Normed U varies between 3217 and 4049, a span of only 800, with a standard deviation of 217.5. The value of 4049 for N = 69 stands out as being very different from the others. If we exclude that one value, the span would only have been 364, with a standard deviation of 115.6. The value 4049 comes from N = 69 and V = 57. For such a small N, index U is very sensitive. If V had been 55 instead of 57, normed U would have changed from 4049 to 3420; i.e. a change of only 2 in V gives a change of more than 600 in norm U for N = 69.

For *SES*, the figures result from table 8 below.

N	*V*	*norm W*	*norm U*	*norm H*
608	348	2465	3151	2762
1212	621	2534	3243	2960
2424	1075	2652	3212	3069
4848	1848	2877	3211	3158
9694	3140	3198	3214	3213
Mean value:		2745	3206	3032
Std. dev :		266.2	30.1	160.1

Table 8. *The same as table 6 except that the text studied is one story from SES.*

Notice in table 8 the extreme constancy of normed U: the difference between the maximum and the minimum value is but 100, although the longest text is 16 times the shortest text.

We see that for the normed indices W and U the tendencies observed for *SNOQ* remain true also for *T* and for *SES*. For the normed index H, however, the situation is completely altered and this index reveals itself not to be very good for these texts. However, in a ln-ln-diagram the figures of N and V still lie on an almost perfect line. Again, regression analysis gives the following line for *T*:

$$\ln V = 0.79583 * \ln N + 0.75529,$$

with a correlation coefficient of 0.9992.

For *SES* we get the line:

$$\ln V = 0.79331 * \ln N + 0.78420.$$

Here the correlation coefficient is 0.9998.

It thus appears that ln-ln regression lines are quite good for deciding if a literary work is homogeneous, but not so for comparing different works with each other.

It is also important to know the 'sensitivity' of an index. By this we understand the effect on the normed index that e.g. a 10% change in V has for different values of N. We therefore

a) increased the observed values of V in table 2 with 10%

and

b) decreased them with the same amount,

and this for different values of N. We found that both norm W and norm H change with 10% as V increases or decreases with 10% and this for all values of N. It can of course be shown mathematically that it always is so. For norm U, on the other hand, we saw that when N is 78, the 10% increase in V results in an increase from 3096 to 4740, which corresponds to 53%. Likewise, the decrease of 10% decreases norm U with 38%, from 3096 to 1933. The situation is reversed when N becomes large. For N = 109000, a 10% change in V affects norm U by only 6%.

A study of the 12 different excerpts of length N = 5000, table 2, shows that the standard deviation in the number of V is 80 words (79.31). The mean value obtained was V = 1675. We can then say that with 95% certainty the actual value of V for N = 5000 lies in the interval $1675 \pm 1.96*80$, i.e. 1675 ± 157, where 1.96 is the usual statistical factor for a 95% confidence interval. Thus the 95% confidence interval is [1518, 1836] and we see that the mean value for norm W of 3007, as in table 6, does not fall within these bounds. For norm U the confidence interval is [2481, 3091], where the mean value of 2846 from table 6 lies well between these bounds. For norm H, we obtain in the same way [2536, 3060]: the mean value of 2550 lies barely over its lower bound. As we can see, in this case only index U is satisfactory. Of course, this discussion can be extended to all different values of N and corresponding confidence intervals for V; however, we have refrained from this.

9. *'Backward counting'.*

Another way to get an understanding of a lexical index is to 'count backwards'. What we mean by this is best explained by an example.

The mean value of V for excerpts from *SNOQ* with N = 1250 (table 2) is 569, which gives W = 9.072, U = 64.61 and H = 4.24.

We now ask: what would the values of V be to give an excerpt of N = 1250 the same values of W, U and H as has the whole of *SNOQ*, namely W = 7.379, U = 64.23 and H = 10.52.

The answer is V = 968 if we use index W, V = 566 if we use index U, and V = 491 for index H (table 9, the line marked with a star); the actual value of V was 569. We see that the excerpt needs almost 400 more forms to get the same index W as has the whole of *SNOQ*, whereas it should have only 3 forms less to get the same index U as the whole of *SNOQ*.

In table 9 below we indicate, for different values of N, first the actual V followed by the standard deviation and then the value of V (called back(W)), that would give the excerpt the same index W as the whole of *SNOQ*. The same is done for index U and finally for index H.

	N	*V*	*Std dev*	*back(W)*	*back(U)*	*back(H)*
	78	60	9	68	58	63
	156	112	7	150	105	105
	313	195	12	301	187	176
	625	356	14	557	328	294
*	1250	569	32	968	566	491
	2500	945	69	1599	964	820
	5000	1675	79	2530	1616	1370
	10000	2796	148	3862	2669	2288
	20000	4560	199	5716	4344	3823
	40000	7332	235	8240	6963	6386
	80000	11180	-	11606	10997	10667
	100000	12854	-	12901	12699	12583
	109598	13466	-	13466	13466	13466

Table 9. *'Backward counts' for SNOQ, according to indices (W), (U) and (H).*

We can see that, except for the case of 40000 words, the difference of the column V and back(U) is less than the standard deviation of V (given in the third column).

Tables 10 and 11 show the corresponding figures for *T* and the chapter from *SES*.

N	*V*	*Std dev*	*back(W)*	*back(U)*	*back(H)*
69	57	4	71	54	80
136	99	3	158	99	132
270	185	4	321	178	219
537	325	10	600	318	364
1072	590	24	1055	561	608
2141	1010	15	1758	980	1014
4280	1783	41	2805	1691	1694
8556	3103	108	4313	2879	2828
17112	5025	329	6427	4843	4726
34225	8273	431	9316	8043	7894
68450	13187	-	13187	13187	13187

Table 10. *'Backward counts' for T, according to indices (W), (U) and (H).*

N	*V*	*Std dev*	*back(W)*	*back(U)*	*back(H)*
608	348	7	451	351	404
1212	645	18	783	617	674
2424	1075	18	1296	1076	1125
4848	1848	17	2054	1849	1880
9694	3140	-	3140	3140	3140

Table 11. *'Backward counts' for SES, according to indices (W), (U) and (H).*

In the three tables above, we can see the very close resemblance between the figures in column back(U) and V, whereas the figures for back(W) are extremely different from V with regards to the standard deviation. Table 10 shows, indeed, that index W becomes absurd in certain situations, e.g. in the first four lines of this table, where back(W) becomes larger than N. In line 4, table 10, index W says that a text of 537 words must have an impossible 600 forms to get the same lexical index as N = 68450 and V = 13187. This again shows that W favours long texts.

A return to table 5 and a scrutiny of the values for index H reveals a marked tendency. The shortest and the longest texts have negative values H, whereas the medium texts have positive values. Also, in diagram 1 the values from *SNOQ* trace a curve which is not exactly straight, but curved slightly concavely downwards. It therefore seems natural to approximate it with a curve of the following kind:

$$\ln V = a*\ln N + b*\ln^2 N + c,$$

where a, b, and c are to be computed by regression analysis, where b<0, because of the concavity.

It is obviously a wish to let the curve pass through the origin, since a one word text (i.e. ln N = 0) always contains exactly one form (i.e. ln V = 0). Therefore we set c = 0.

By regression analysis we get the following values from *SNOQ*: a = 1.0144, and b = -0.01675. From *T* we get a = 1.00946 and b = -0.01392. We see that in both cases a is close to 1. In order to operate with just one number, namely b, and not two numbers, a and b, we therefore put a = 1 and compute only b. A computation then gives

$$-1/b = \ln^2 N/(\ln N - \ln V),$$

where we recognize - 1/b as index U in formula (2).[12]

[12] In Brunet (1985, vol 1:33) a regression analysis for 22 texts by Zola gives a formula where the constant 'a' (in Brunet called 'b') is equal to 1.006 .

Thus index U is the natural index if we want a regression curve of degree 2 in ln V and ln N which passes through the origin, i.e. if we postulate that a one-word text must have exactly one word form, which we believe to be a quite reasonable assumption.

10. Non-homogeneous works.

We also did some experimenting on other texts by Primo Levi, namely *Racconti e saggi* (hence referred to as *RS*), *L'altrui mestiere* (hence referred to as *AM*), *Vizio di forma* (hence referred to as *VF*), *Storie naturali* (hence referred to as *SN*) and *Sistema periodico* (hence referred to as *SP*). All of these works are collections of essays, short stories and small articles from a newspaper column, all on different subjects. We measured the normed index for each essay or story separately, took the mean value for the stories in each book and the standard variation, and compared the result with the normed index for the whole book, considered as an entity. We expected that the normed index of the whole book would be greater than the normed indices from the separate chapters or stories. The results we obtained are shown in table 12 below.

Text	*Mean W*	*std dev*	*Whole W*	*Mean U*	*std dev*	*Whole U*	*Mean H*	*std dev*	*Whole H*
RS	2472	115	4425	3125	255	3386	2863	148	3420
AM	2547	95	4941	3232	200	3318	2967	126	3283
VF	2485	166	4160	2905	274	3034	2832	174	2908
SN	2661	159	4245	3124	298	3130	3003	161	3053
SP	2688	268	4744	3119	338	3227	3036	271	3152

Table 12. Mean values of parts and whole of other texts by Levi, computed according to indices W, U, and H.

It can be noticed that the difference between the mean value for the separate stories and the whole book as an entity is very large when index W is used and astonishingly small when we use index U. In the case of *SN* and index U the difference is practically nil. This may be a point of further study.

We also measured the two stories of *RS* which scored lowest in norm U, namely *La grande mutazione* (N = 1442, V = 642) and *Nozze della formica* (N = 1205, V = 543). According to U, they received 2734 and 2619 respectively. When we combined them into one text file, this file

measured 2823 in norm U. The fact that the combined text file of two stories scores a higher normed (U) is, of course, due to the fact that each of the two separate stories has its own setting and thus its own vocabulary; thus the combined text file was not considered a homogeneous text. We carried out several other experiments of this kind, and they all showed the same tendency.

In fact, we expected the differences in values between what we here considered homogeneous and non-homogeneous texts to be greater. In view of these revealed facts, texts D and E accounted for above might even be lexically more homogeneous than we would have believed them to be, were they written by the same author. In this context we should like to cite Brunet (1988, vol 1:39):

> "Les monographies d'auteurs constituent une unité organique où les textes ne sont pas pleinements cumulatifs: même si les sujets ou les genres diffèrent, les productions d'un même écrivain ont une zone lexicale comune, et la part privative (c'est à dire la fréquence 1) tend à être plus étroite. Le corpus du TLF au contraire est fait de l'assemblage d'écrivains divers, de la superposition d'époques et de genres différents et la fréquence 1 témoigne de cette disparate."[13]

11. Conclusions.

We have here discussed some ways to compute and thus measure the lexical richness of a text, using various indices. We have studied the indices W, U, and H primarily on literary texts that we have considered to be homogeneous in the sense given in the postulates. We have found that index U seems to be more impartial to the length of a text than the other two indices.

We have also discussed the possibilities of interpreting the results of such computing in more tangible terms than those of the traditional abstract figures which are given by conventional indices. With these tangible terms we hope to increase the comprehension of how a lexical index works. We also hope to have shown the differences between the various indices discussed. From our experiments we have only been able to draw conclusions about how the indices work on Primo Levi's texts; it is also our intention to carry out similar experiments on non-literary texts. As a conclusion from the computing on indices W, U, and H described above, I have decided only to work with index U in my further studies. For future investigations, however, I intend to continue my experiment on other literary as well as non-literary texts. It is my sincere wish that other scholars of stylolinguistics will accept the challenge to engage in similar studies with their 'own' corpora, so that

[13] TLF = *Trésor de la Langue Française*, one of the main objects of Brunet's studies.

the various indices will be tried out further on different authors of different languages and genres.

Address of the Author:

Department of French and Italian
Stockholm University
S - 10691 Stockholm
Sweden

References

Bernet, Ch. (1983), *Le vocabulaire des tragédies de Jean Racine. Analyse statistique*, Genève-Paris, Slatkine-Champion.

Brunet, E. (1978), *Le vocabulaire de Giraudoux. Structure et évolution*, Genève-Paris, Slatkine-Champion.

Brunet, E. (1983), *Le vocabulaire de Proust. Étude quantitative*, Vols 1-3, Genève-Paris, Slatkine-Champion.

Brunet, E. (1985), *Le vocabulaire de Zola. Étude quantitative*, Vols 1-3, Genève-Paris, Slatkine-Champion.

Brunet, E. (1988), *Le vocabulaire de Victor Hugo*, Vols 1-3, Genève-Paris, Slatkine-Champion.

Dugast. D. (1978), "Sur quoi se fond la notion d'étendue théorique du vocabulaire", *Le Francais Moderne* 46: 25-32.

Dugast, D. (1979), *Vocabulaire et discours. Essais de lexicométrie organisationnelle*, Genève, Slatkine.

Efron, B. & Thisted, R. (1976), "Estimating the number of unseen species: How many words did Shakespeare know?", *Biometrika*, 63: 435-47.

Guiraud, P. (1954), *Les caractères statistique du vocabulaire*, Paris, Puf.

Guiraud, P. (1959), *Problèmes et méthodes de la statistique linguistique*, Dortrecht, Reidel Publishing Company.

Herdan, G. (1966), *The advanced theory of language as choice and chance*, Berlin-Heidelberg-New York, Springer.

Levi, P. (1987-1990), *Opere*, 1-3, Torino, Einaudi.

Muller, Ch. (1968), *Initiation à la statistique linguistique*, Paris, Larousse.

Muller, Ch. (1971), "Sur la mesure de la richesse lexicale. Théorie et expériences", *Études de linguistique appliquées*, Jan-mar 1971; nouv. ser. 1: 20-46.

Muller, Ch. (1973), *Initiation aux méthodes de la statistique linguistique*, Paris, Hachette.

Muller, Ch. (1977), *Principes et méthodes de la statistique lexicale*, Paris, Hachette.

Nystedt, J. (1993), *Le opere di Primo Levi viste al computer. Osservazioni stilolinguistiche*, Acta Universitatis Stockholmiensis. Romanica Stockholmiensia 14, Stockholm, Almqvist & Wiksell.

Picchi, E. (1989), *DBT: Data Base Testuale*, ILC-DBT-1, Pisa.